DOCUMENT RESUME

ED 410 280 TM 027 091

AUTHOR Meshbane, Alice; Morris, John D.
TITLE A Method for Selecting between Fisher's Linear Classification Functions and Least Absolute Deviation in Predictive Discriminant Analysis.
PUB DATE Mar 97
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classification; *Comparative Analysis; Computer Software; Correlation; *Discriminant Analysis; *Prediction
IDENTIFIERS *Cross Validation; *Fishers Linear Classification Functions

ABSTRACT

A method for comparing the cross-validated classification accuracy of Fisher's linear classification functions (FLCFs) and the least absolute deviation is presented under varying data conditions for the two-group classification problem. With this method, separate-group as well as total-sample proportions of current classifications can be compared for the two classification procedures. Q. McNemar's (1947) test for contrasting correlated proportions is used in statistical comparisons of separate-group and total-sample proportions. The method is illustrated with 22 real data sets. FLCFs were built on assumptions of multivariate normality, equal covariance matrices, and equal prior probabilities of group membership. Least absolute deviation (LAD) models were built using a computer subroutine that incorporates linear programming code. Use of the method and computer program demonstrated in this study will allow researchers to compare the explicit cross-validated classification hit-rate accuracy of LAD and FLCFs for any specific data set and select the procedure that yields the higher total-sample or separate-group hit rate, depending on the hit rate of interest. (Contains 1 table and 23 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 410 280

A Method for Selecting Between Fisher's Linear Classification Functions and Least Absolute Deviation in Predictive Discriminant Analysis

Alice Meshbane and John D. Morris
Florida Atlantic University

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Alice Meshbane
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the American Educational Research Association, March 1997, Chicago, IL

TM027091

BEST COPY AVAILABLE

A Method for Selecting Between Fisher's Linear Classification Functions and Least Absolute Deviation in Predictive Discriminant Analysis

ABSTRACT. A method for comparing the cross-validated classification accuracy of Fisher's linear classification functions and least absolute deviation is presented under varying data conditions for the two-group classification problem. With this method, separate-group as well as total-sample proportions of correct classifications can be compared for the two classification procedures. McNemar's (1947) test for contrasting correlated proportions is used in statistical comparisons of separate-group and total-sample proportions. The method is illustrated with 22 real data sets.

Standard techniques for solving the two-group classification problem, such as ordinary least squares (OLS) and Fisher's linear classification functions (FLCFs)[1], are based on assumptions of multivariate normality and equality of group covariance matrices. Mathematical programming (MP) solution techniques are not. Consequently, we might expect an MP approach to yield greater cross-validated classification accuracy than a standard approach for nonnormal data or data with unequal group covariance matrices.

Joachimsthaler and Stam (1990) reviewed the literature on studies comparing the accuracy of MP and standard approaches, and concluded that some MP approaches outperform some standard approaches some of the time, but no MP approach outperforms all standard approaches all of the time. Thus, studies comparing the cross-validated classification accuracy of MP and standard approaches (e.g., Bajgier & Hill, 1982; Freed & Glover, 1986; Joachimsthaler & Stam, 1990; Stam & Joachimsthaler, 1989, 1990) provide only partial support for the expectation that MP approaches will be more accurate than standard approaches when assumptions are violated.

One MP technique that shows promise as an alternative to standard classification approaches is least absolute deviation (LAD). This procedure was discussed by Armstrong, Frome, and Sklar (1980) in the multiple regression context as an alternative to OLS, especially for applications involving nonnormal distributions. Morris and Huberty (1983) surmised that LAD might also be appropriate for classification because the absolute rather than squared deviation between predicted and actual dichotomous scores determines the group to which a case is classified.

This conjecture has received empirical support from two simulation studies. Morris and Huberty (1983) compared the cross-validated classification accuracy of

LAD and OLS across a wide variety of data conditions, and found a small but consistent advantage of LAD over OLS. Lee and Ord (1990) compared LAD to several MP solution techniques and to OLS, and concluded that LAD performs at least as well as these other procedures. Furthermore, Lee and Ord note that LAD is easy to implement and inexpensive to run.

At this point, then, we know that LAD has performed as well as, or better than, OLS using simulated data. However, we do not know whether this result replicates when LAD is compared with FLCFs using real data sets. To address this issue, a method for comparing the cross-validated classification accuracy of LAD and FLCFs for a specific data set is introduced and demonstrated. This method will enable researchers to select the optimal classification procedure for a specific data set, regardless of data conditions.

Method

Data Source: Initially, 33 classification data sets varying in number of cases, number of predictor variables, degree of group separation, and equality of group covariance matrices were selected to illustrate the method. However, for 11 of these data sets, the algorithm used to obtain weights for LAD failed to converge. Consequently, only 22 data sets were employed to illustrate the method. To bolster validity, all of the data sets were taken from real classification studies (Hekelman, Zyzanski, & Flocke, 1995; Morris & Huberty, 1987).

Procedure: FLCFs were built based on assumptions of multivariate normality, equal covariance matrices, and equal prior probabilities of group membership. Cases were classified into groups using Tatsuoka's (1988, p. 351) minimum chi-squared rule. We assumed equal priors because population sizes were unknown. The use of sample sizes as estimates of relative population sizes is not recommended when population sizes are unknown (Huberty, 1994, p. 65; Meshbane & Morris, 1995b).

LAD models were built using a computer subroutine that incorporates linear programming code (Armstrong, Frome, & Kung, 1979). The weights generated by this algorithm minimize the sum of absolute deviations between predicted and actual criterion scores.

In comparing the predictive accuracy of LAD and FLCFs, external rather than internal results were considered. Results of an internal classification analysis are those obtained when measures for the individuals on whom the statistics were based are resubstituted to obtain the predicted classification scores. In an external classification

analysis, statistics based on one set of individuals are used in classifying new individuals. An external analysis is appropriate for making inferences about the discriminatory power of the predictors for a new set of data (Huberty, 1994, p. 110).

External, or cross-validated, hit-rate accuracy was estimated using the leave-one-out procedure. A case was classified by applying the function derived from all cases except the one being classified. This process was repeated round-robin for each case with a count of the overall classification accuracy used to estimate the cross-validated accuracy. This procedure, and the analogous PRESS (PRedicted Error Sum of Squares) procedure (Allen, 1971) for multiple regression, has a relatively wide following in both the discriminant analysis and multiple regression literature (see, for example, Allen, 1971; Allen & Cady, 1982; Huberty, 1994; Huberty & Mourad, 1980; Lachenbruch, 1967; Mosteller & Tukey, 1968).

McNemar's (1947) statistic for correlated proportions was used in the statistical comparisons of LAD and FLCF hit rates for separate-group and total-sample proportions. This method was previously suggested for comparing full and reduced classification models (Morris & Huberty, 1995) and linear and quadratic classification models (Meshbane & Morris, 1995a), and for selecting predictor variable subsets (Morris & Meshbane, 1995), but is equally applicable in comparing LAD and FLCFs. (See Looney, 1988, for a method of comparing classification results of more than two models.) Because the calculation of the McNemar correlated proportion statistic requires the joint distribution of hits and misses for both LAD and FLCFs, no statistical package will accomplish the method. Therefore, we wrote a FORTRAN computer program to provide this information, as well as to perform the other functions described in this section (i.e., determine LAD and FLCF weights, classify cases into groups, conduct the McNemar test and the Box test).

We used the Box test for testing the assumption of homogeneity of covariance structures. This test is sensitive to multivariate normality, and the outcome is therefore confounded with the homogeneity of dispersion issue. Nevertheless, the Box test is routinely used for testing the homogeneity of dispersion assumption and is even the default in some statistical packages. Notwithstanding concerns over this test, one could argue that, theoretically, the LAD procedure is more likely to be appropriate when the Box test indicates that the covariance structures are unequal.

Results

For each of the data sets, Table 1 gives a short description, the degree of group separation (D) calculated using the centroid to centroid formula described by Huberty

(1994, p. 43), an index of disproportionality of the group sizes (I) calculated as (n_i * 100) / n_j, where n_i is the larger of the two groups, the number of cases in group 1 (n_1), the number of cases in group 2 (n_2), the number of predictor variables (p), results of the Box test for homogeneity of covariance structures, and a comparison of the leave-one-out (L-O-O) performance of LAD and FLCFs for the total sample and separately for each group. We compared the performance of the two classification procedures, displayed as the hit-rate percent obtained by the p predictor variables, via McNemar's test for contrasting correlated proportions.

Insert Table 1 about here

To illustrate the method for these data sets, we used the .01 alpha level with the associated critical z of 2.58. As can be seen in Table 1, differences between LAD and FLCFs in classifying the total sample were statistically significant in only two of the data sets (15 and 22). In both cases, FLCFs yielded significantly higher hit rates. Differences between the two classification procedures in separate-group hit rates were statistically significant in seven of the 22 data sets (8, 9, 10, 11, 15, 21, 22). In five of these data sets (10, 11, 15, 21, 22), FLCFs worked better. In the other two data sets (8, 9), LAD worked better.

To assess practical significance, we calculated an index of improvement in hit rate over chance (described by Huberty, 1994, p. 107) for both LAD and FLCFs. We defined chance using the proportional chance criterion (see Huberty, 1994, p. 103). Our effect size measure was the difference between LAD and FLCFs in improvement over chance. Using a critical effect size of 10%, all statistically significant results were also of practical significance (effect sizes ranged from 15.84% in data set 11 to 47.25% in data set 8).

Discussion

While these results are partially consistent with those of Lee and Ord (1990), who reported similar misclassification rates for LAD and OLS procedures, the results are clearly inconsistent with Morris and Huberty (1983), who reported a small but consistent advantage of LAD over OLS across all data sets. The discrepancy between these findings and those of the current study may be related to relative group size. Specifically, in both the Lee and Ord and the Morris and Huberty studies, the sizes of the two groups were identical (Lee and Ord) or nearly so (Morris and Huberty). In the

current study, the sizes of the two groups were different (index of disproportionality 118 or higher) in eight of the 22 data sets.

For relative group size to be a plausible explanation for the inconsistent results, we would expect data sets with statistically significant differences in performance for LAD and FLCFs to have non-trivial differences in group sizes, and data sets yielding similar performance for the two procedures to have similar group sizes. As evident from Table 1, group sizes were different (index of disproportionality was 118 or higher) in five of the seven data sets with statistically significant differences in separate group hit rates between LAD and FLCFs. Furthermore, relative group sizes were similar (index of disproportionality was 113 or lower) in 12 of the 15 data sets in which the performance of LAD and FLCFs were similar. In the other three data sets, the Mahalanobis distance between groups was relatively high (D = 2.89 to 3.11), which may have negated the conjectured effect of discrepant group sizes in these data sets.

Regardless of the reason(s) for the discrepant results, the relationship between data conditions and model performance is far from perfect. Thus, for a given data set, it is not yet possible to predict the relative performance of LAD and FLCFs on the basis of data conditions alone. Use of the method and computer program[2] demonstrated herein will allow researchers to compare the explicit cross-validated classification hit-rate accuracy of LAD and FLCFs for any specific data set and select the procedure that yields the higher total-sample or separate-group hit rate, depending on which hit rate is of interest.

Notes

1. A linear classification function is different from a linear discriminant function.

2. For a copy of the FORTRAN program that accomplishes the method, send a returnable diskette and diskette mailer to Alice Meshbane, College of Education, Florida Atlantic University, P.O. Box 3091, Boca Raton, FL 33431-0991. Internet: Meshbane@acc.fau.edu.

References

Allen, D. M. (1971). The prediction sum of squares as a criterion for selecting predictor variables (Tech. Rept. No. 23). University of Kentucky, Department of Statistics.

Allen, D. M., & Cady, F. B. (1982). Analyzing experimental data by regression. Belmont, CA: Wadsworth.

Armstrong, R. D., Frome, E. L., & Kung, D. S. (1979). A revised simplex algorithm for the absolute deviation curve fitting problem. Communications in Statistics: Simulation and Computation, B8, 175-190.

Armstrong, R. D., Frome, E. L., & Sklar, M. G. (1980). Linear programming in exploratory data analysis. Journal of Educational Statistics, 4, 293-307.

Bajgier, S. M., & Hill, A. V. (1982). An experimental comparison of statistical and linear programming approaches to the discriminant problem. Decision Sciences, 13, 604-618.

Freed, N., & Glover, F. (1986). Evaluating alternative linear programming models to solve the two-group discriminant problem. Decision Sciences, 17, 151-162.

Hekelman, F. P., Zyzanski, S. J., & Flocke, S. A. (1995). Research in Higher Education, 36, 235-255.

Huberty, C. J. (1994). Applied discriminant analysis. New York: Wiley.

Huberty, C. J, & Mourad, S. A. (1980). Estimation in multiple correlation/prediction. Educational and Psychological Measurement, 40, 101-112.

Joachimsthaler, E. A., & Stam, A. (1990). Mathematical programming approaches for the classification problem in two-group discriminant analysis. Multivariate Behavioral Research, 25, 427-454.

Lachenbruch, P. A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. Biometrics, 23, 639-645.

Lee, C. K., & Ord, J. K. (1990). Discriminant analysis using least absolute deviations. Decision Sciences, 21, 86-96.

Looney, S. W. (1988). A statistical technique for comparing the accuracies of several classifiers. Pattern Recognition Letters, 8, 5-9.

McNemar, Q. (1947). Note on the sampling error of the differences between correlated proportions or percentages. Psychometrika, 12, 153-157.

Meshbane, A., & Morris, J. D. (1995a). A method for selecting between linear and quadratic classification models in discriminant analysis. The Journal of Experimental Education, 63, 263-273.

Meshbane, A., & Morris, J. D. (1995b, April). Assuming equal vs. unequal prior probabilities of group membership in discriminant analysis: Effect on predictive accuracy. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Morris, J. D., & Huberty, C. J (1983, April). Some alternative classification weighting algorithms. Paper presented at the annual convention of the American Educational Research Association, Montreal.

Morris, J. D. & Huberty, C. J (1995). Full versus restricted model testing in discriminant analysis. Journal of Experimental Education, 63, 161-165.

Morris, J. D., & Meshbane, A. (1995). Selecting predictor variables in two-group classification problems. Educational and Psychological Measurement, 55, 438-441.

Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), Handbook of Social Psychology: Vol. 2. (pp. 80-203). Reading, MA: Addison-Wesley.

Stam, A., & Joachimsthaler, E. A. (1989). Solving the classification problem in discriminant analysis via linear and nonlinear programming methods. Decision Sciences, 20, 285-293.

Stam, A., & Joachimsthaler, E. A. (1990). A comparison of a robust mixed-integer approach to existing methods for establishing classification rules for the discriminant problem. European Journal of Operations Research, 46, 113-132.

Tatsuoka, M. J. (1988). Multivariate analysis: Techniques for educational and psychological research (2nd ed.). New York: Macmillan.

Table 1

Data Set Description, Results of Box Test for Equality of Covariance Matrices, and Comparison of Hit Rate Percents for LAD and FLCFs Assuming Equal Population Sizes

#	Data Set Description	D	I	n_1	n_2	p	Results of Box Test for Equal Covariance Matrices	Procedure Used	L-O-O Hit Rate % Total	GR 1	GR 2
1	Fisher Data - Groups 1 & 3	13.97	100	50	50	4	$\chi^2 = 6.9057$, $p < .0001$	FLCFs	100	100	100
							df = 10	LAD	100	100	100
								McNemar's z	.00	.00	.00
2	Fisher Data - Groups 1 & 2	10.16	100	50	50	4	$\chi^2 = 5.0455$, $p < .0001$	FLCFs	100	100	100
							df = 10	LAD	100	100	100
								McNemar's z	.00	.00	.00
3	Bisbey Data - Groups 1 & 3	5.12	106	35	37	13	$\chi^2 = .9939$, $p = .5013$	FLCFs	97	94	100
							df = 91	LAD	96	91	100
								McNemar's z	1.00	1.00	.00
4	Fisher Data - Groups 2 & 3	3.77	100	50	50	4	$\chi^2 = .7148$, $p = .7125$	FLCFs	97	96	98
							df = 10	LAD	96	94	98
								McNemar's z	.58	.58	.00

Table 1 cont.

Data Set Description, Results of Box Test for Equality of Covariance Matrices, and Comparison of Hit Rate Percents for LAD and FLCFs Assuming Equal Population Sizes

#	Data Set Description	D	I	n_1	n_2	p	Results of Box Test for Equal Covariance Matrices	Procedure Used	L-O-O Hit Rate % Total	GR 1	GR 2
5	Junior Faculty Performance Data	3.11	152	27	41	4	$\chi^2 = 2.2172$, $p = .0150$	FLCFs	91	93	90
							$df = 10$	LAD	87	85	88
								McNemar's z	1.73	1.41	1.00
6	Rulon Data - Groups 1 & 3	2.93	129	85	66	4	$\chi^2 = 3.4973$, $p = .0003$	FLCFs	93	94	91
							$df = 10$	LAD	93	94	91
								McNemar's z	.00	.00	.00
7	Bisbey Data - Groups 1 & 2	2.89	231	35	81	13	$\chi^2 = 1.0021$, $p = .4777$	FLCFs	89	89	89
							$df = 91$	LAD	87	74	93
								McNemar's z	.71	2.24	-1.73
8	Bisbey Data - Groups 2 & 3	2.41	219	81	37	13	$\chi^2 = 1.2131$, $p = .0929$	FLCFs	84	83	87
							$df = 91$	LAD	91	98	76
								McNemar's z	-1.89	-3.21	2.00

Table 1 cont.

Data Set Description, Results of Box Test for Equality of Covariance Matrices, and Comparison of Hit Rate Percents for LAD and FLCFs Assuming Equal Population Sizes

#	Data Set Description	D	I	n_1	n_2	p	Results of Box Test for Equal Covariance Matrices	Procedure Used	L-O-O Hit Rate % Total	GR 1	GR 2
9	Talent Data - Groups 3 & 5	1.97	285	33	94	14	$\chi^2 = 1.1086$, $p = .2238$	FLCFs	77	79	77
							$df = 105$	LAD	83	67	88
								McNemar's z	-1.81	2.00	-3.32
10	Demographic # 2 - Body Char	1.88	129	157	122	8	$\chi^2 = 6.6870$, $p < .0001$	FLCFs	82	83	82
							$df = 36$	LAD	79	85	72
								McNemar's z	2.32	-1.73	3.46
11	Rulon Data - Groups 2 & 3	1.87	141	93	66	4	$\chi^2 = 3.4962$, $p = .0003$	FLCFs	83	85	80
							$df = 10$	LAD	82	91	70
								McNemar's z	.28	-2.45	2.65
12	Rulon Data - Groups 1 & 2	1.74	109	85	93	4	$\chi^2 = 4.9003$, $p < .0001$	FLCFs	81	84	79
							$df = 10$	LAD	82	81	82
								McNemar's z	-.45	1.41	-1.73

Table 1, cont.

Data Set Description, Results of Box Test for Equality of Covariance Matrices, and Comparison of Hit Rate Percents for LAD and FLCFs Assuming Equal Population Sizes

#	Data Set Description	D	I	n_1	n_2	p	Results of Box Test for Equal Covariance Matrices	Procedure Used	L-O-O Hit Rate % Total	GR 1	GR 2
13	Talent Data - Groups 1 & 5	1.72	113	83	94	14	$\chi^2 = 1.5086$, $p = .0014$	FLCFs	75	74	76
							$df = 105$	LAD	78	77	79
								McNemar's z	-1.50	-1.00	-1.13
14	Demographic # 3 - Body Char	1.36	104	142	137	8	$\chi^2 = 5.2724$, $p < .0001$	FLCFs	73	70	76
							$df = 36$	LAD	73	68	79
								McNemar's z	.00	1.63	-1.41
15	Block Data - Groups 3 & 4	.85	100	38	38	4	$\chi^2 = 4.3098$, $p < .0001$	FLCFs	68	74	63
							$df = 10$	LAD	43	58	29
								McNemar's z	3.96	1.90	3.61
16	Block Data - Groups 1 & 2	.84	105	40	38	4	$\chi^2 = 2.2028$, $p = .0157$	FLCFs	68	60	76
							$df = 10$	LAD	63	48	79
								McNemar's z	1.63	2.24	-1.00

Table 1, cont.

Data Set Description, Results of Box Test for Equality of Covariance Matrices, and Comparison of Hit Rate Percents for LAD and FLCFs Assuming Equal Population Sizes

#	Data Set Description	D	I	n_1	n_2	p	Results of Box Test for Equal Covariance Matrices	Procedure Used	L-O-O Hit Rate % Total	GR 1	GR 2
17	Block Data - Groups 1 & 4	.81	105	40	38	4	$\chi^2 = 1.5500$, $p = .1163$	FLCFs	60	58	63
							$df = 10$	LAD	59	53	66
								McNemar's z	.38	1.00	-.58
18	Block Data - Groups 1 & 3	.74	103	40	39	4	$\chi^2 = 5.3857$, $p < .0001$	FLCFs	65	58	72
							$df = 10$	LAD	57	45	69
								McNemar's z	1.90	2.24	.45
19	Block Data - Groups 2 & 3	.64	105	37	39	4	$\chi^2 = 4.5542$, $p < .0001$	FLCFs	55	57	54
							$df = 10$	LAD	49	54	44
								McNemar's z	1.29	.33	1.63
20	Block Data - Groups 2 & 4	.52	103	37	38	4	$\chi^2 = 1.2033$, $p = .2838$	FLCFs	59	62	55
							$df = 10$	LAD	61	65	58
								McNemar's z	-1.42	-1.00	-1.00

Table 1, cont.

Data Set Description, Results of Box Test for Equality of Covariance Matrices, and Comparison of Hit Rate Percents for LAD and FLCFs Assuming Equal Population Sizes

#	Data Set Description	D	I	n_1	n_2	p	Results of Box Test for Equal Covariance Matrices	Procedure Used	L-O-O Hit Rate % Total	GR 1	GR 2
21	Demographic # 1 - Body Char	.50	104	137	142	8	$\chi^2 = 5.4808$, $p < .0001$	FLCFs	58	61	55
							$df = 36$	LAD	55	53	57
								McNemar's z	1.96	3.46	-1.00
22	Warncke Data - Groups 2 & 3	.45	118	47	40	10	$\chi^2 = 1.2556$, $p = .1039$	FLCFs	40	45	35
							$df = 55$	LAD	28	43	10
								McNemar's z	2.84	.45	3.16